Capstone Project-2

Yes Bank Stock Closing Price Prediction

Team Members

Aditya Singh Thakur
Meenakshi
Tushar R. Wagh

Let's Predict!

1. Overview & Objective
2. Data Pipeline
3. EDA
4. Regression Analysis
    a) Linear Regression
    b) Ridge Regression with and without Cross Validation
    c) Lasso Regression with and without Cross Validation
    d) Elastic Net Regression with and without CV
    e) KNeighbor Regressor
    f) Support Vector Regression
5. Conclusion

Overview & Objective

- **Overview**

- Yes Bank is a well-known bank in the Indian financial domain. Since 2018, it has been in the news because of the fraud case involving Rana Kapoor. Owing to this fact, it was interesting to see how that impacted the stock prices of the company and whether any predictive models can do justice to such situations..

Objective

This dataset has monthly stock prices of the bank since its inception and includes closing, starting, highest, and lowest stock prices of every month. The main objective is to predict the stock's closing price of the month.

Data Pipeline

Data Preprocessing: At this stage,

- we check for duplicate values and missing values and treat them if any.
- Detecting the outliers and removed it.
- we check the datatype of the features present in our dataset
  transform them if necessary.

Exploratory Data Analysis (EDA): At this stage, we conduct an EDA on the selected features in order to better understand their spread , pattern and relationship with the other features. It gives us an intuition as to what is going on in the dataset.

Model Building: At this stage, we apply various models to understand which one will give us the best result.

Data Summary

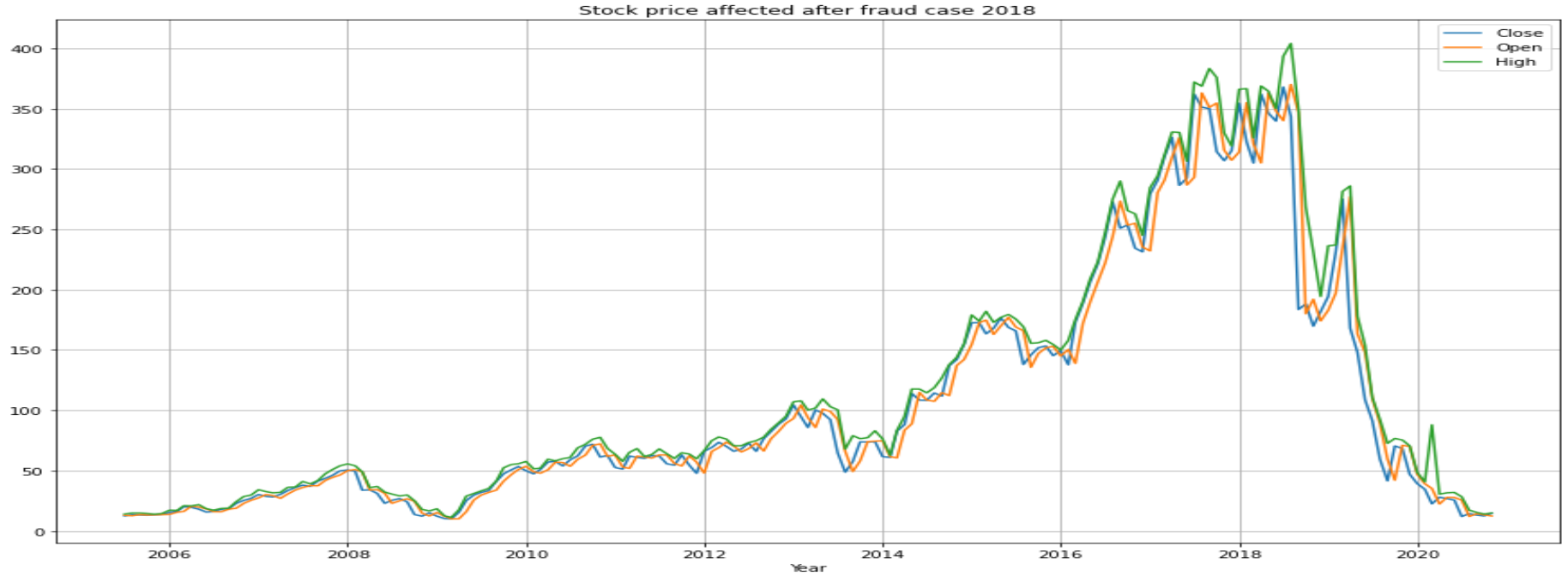
We have Yes Bank monthly stock price dataset. It has following features (Columns):

1) **Open** : Opening price of the stock of particular day
2) **High** : It's the highest price at which a stock traded during a period
3) **Low** : It’s the lowest price at which stock traded during a period
4) **Close** : Closing price of a stock at the end of a Trading Day
5) **Date** : We will use it as a index

Note: ‘Close’ will be our Dependent variable & Others will be independent.

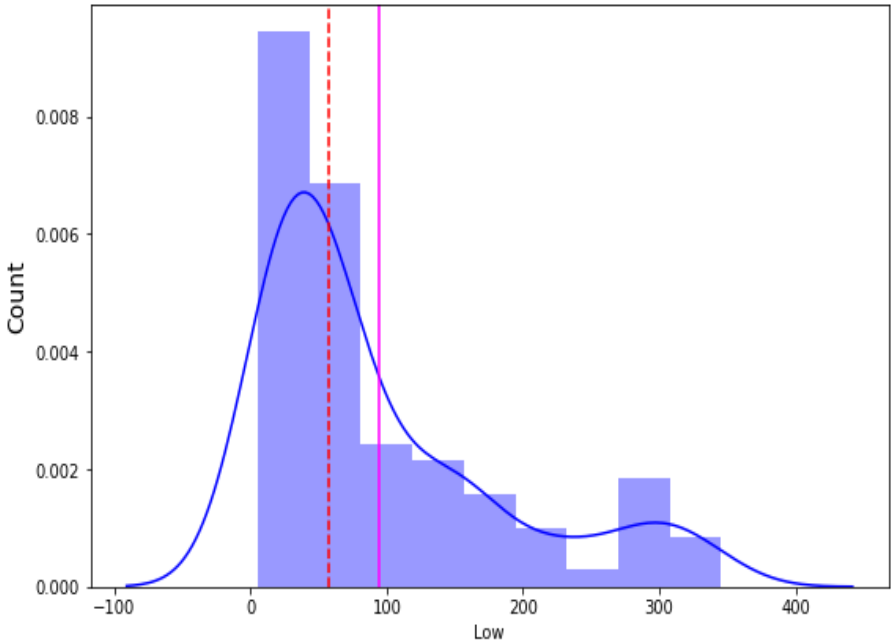
Exploratory Data Analysis(EDA)

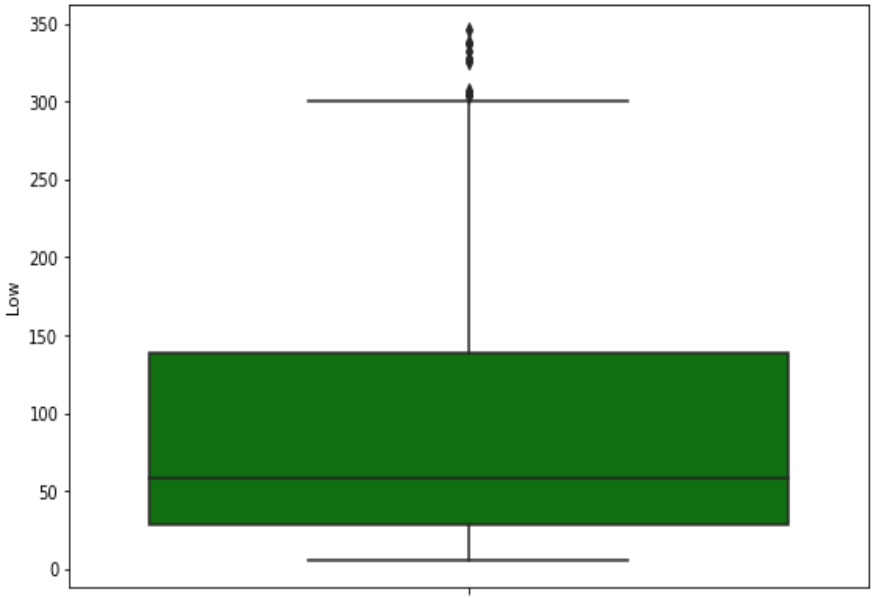
Let plot each numerical column and see how stock price is affected after fraud case 2018



- From the above line plot ,We conclude that the stock price is keep on increasing till 2018.
- But after 2018 , the stock price is keep on decreasing due the fraud case involving Rana Kapoor.

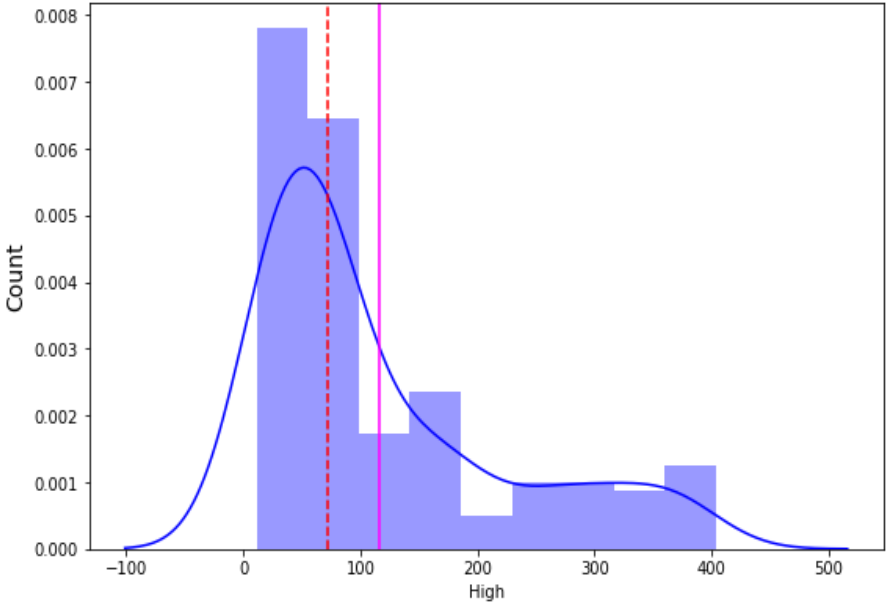
Distribution of ‘Low’ Stock Price

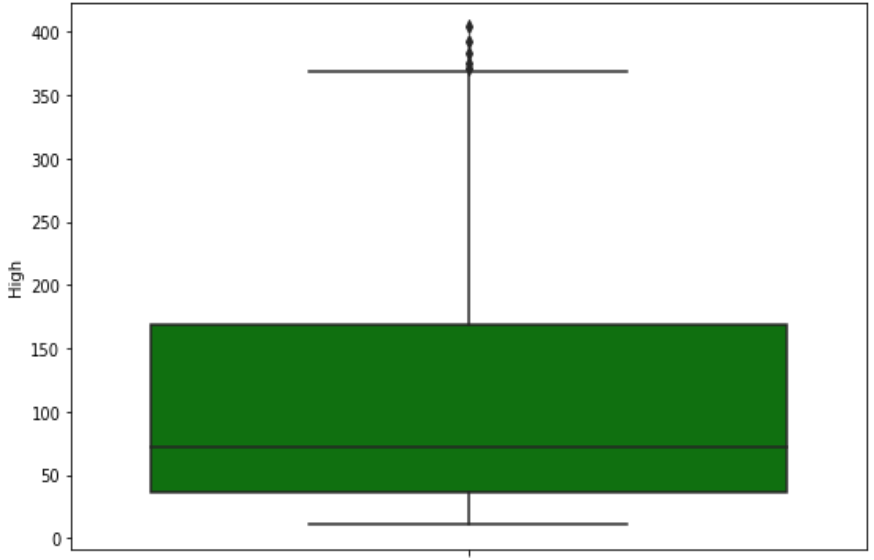




- The above graph shows that it is not a normal distribution curve.
- The mean and median should be equal for perfect normal distribution curve.But,mean is not equal to median as there is not a perfect normal distribution curve.
- We need to convert all the features to normal distribution using log transformation.
- Outliers are present in each column.By, converting our features to normal distribution using log transform .We can remove outliers from the dataset.

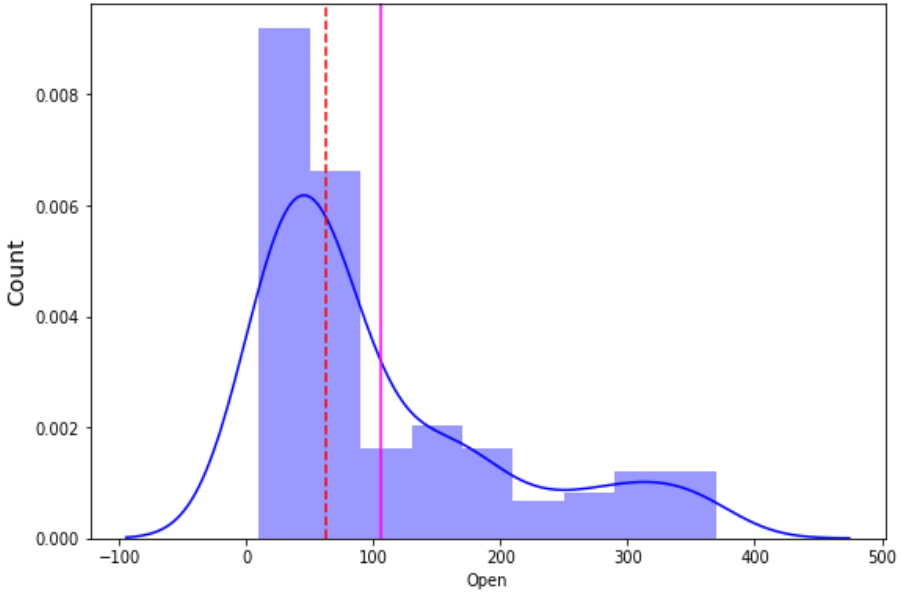
Distribution of ‘High’ Stock Price

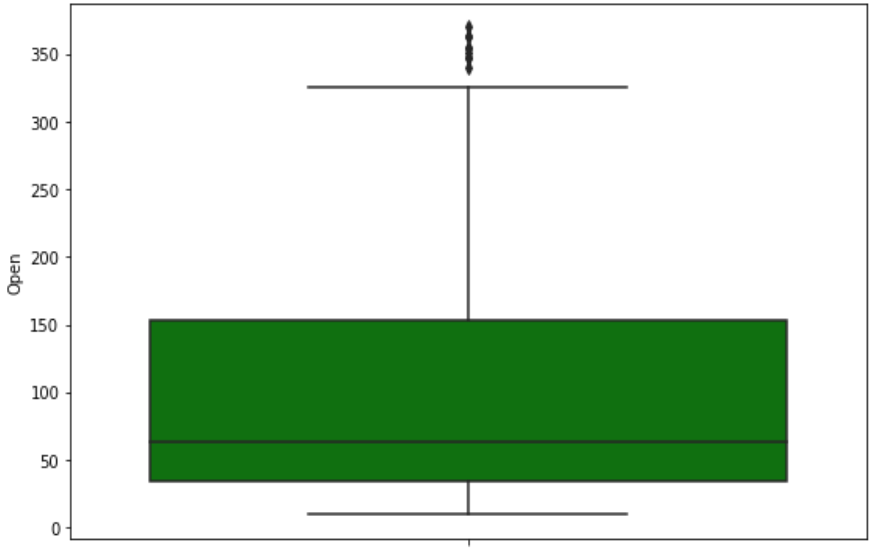




- The above graph shows that it is not a normal distribution curve.
- The mean and median should be equal for perfect normal distribution curve.But,mean is not equal to median as there is not a perfect normal distribution curve.
- We need to convert all the features to normal distribution using log transformation.
- Outliers are present in each column.By,converting our features to normal distribution using log transform .We can remove outliers from the dataset.

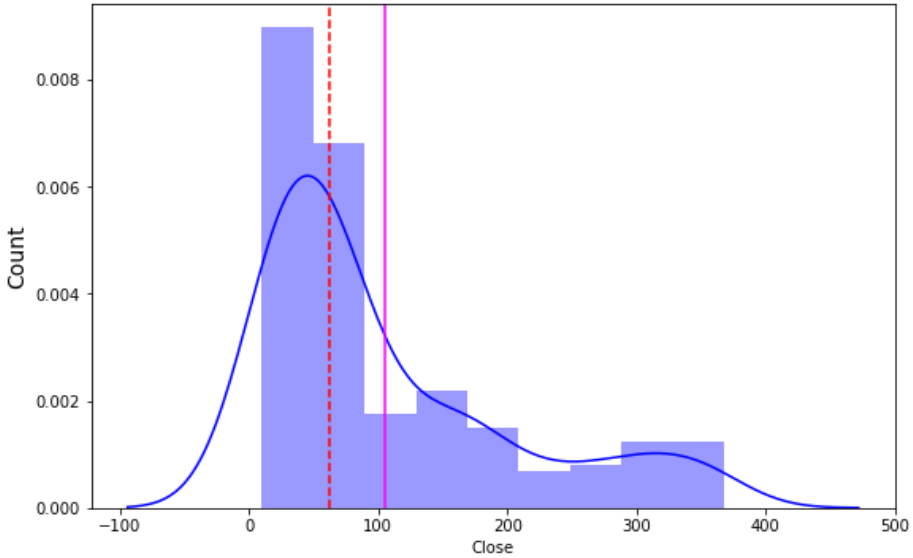
Distribution of ‘Open’ Stock Price

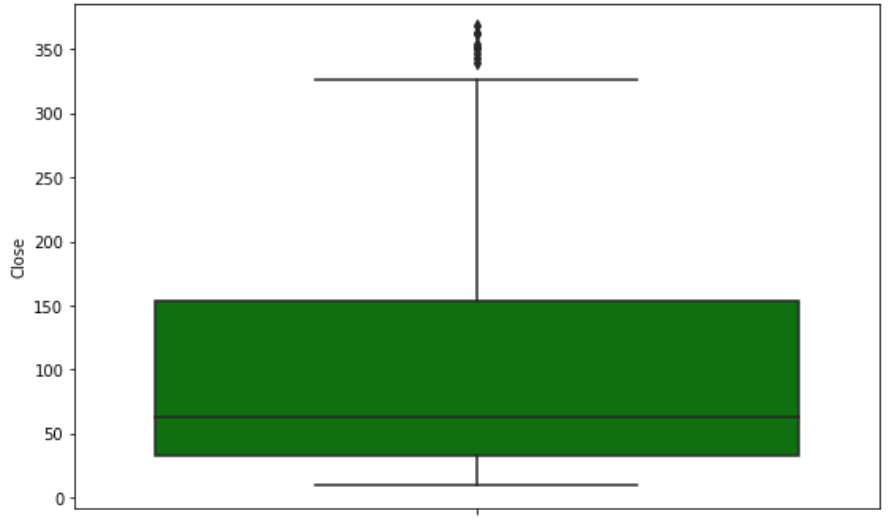




- The above graph shows that it is not a normal distribution curve.
- The mean and median should be equal for perfect normal distribution curve.But,mean is not equal to median as there is not a perfect normal distribution curve.
- We need to convert all the features to normal distribution using log transformation.
- Outliers are present in each column.By,converting our features to normal distribution using log transform .We can remove outliers from the dataset.

Distribution of ‘Close’ Stock Price(Dependent)





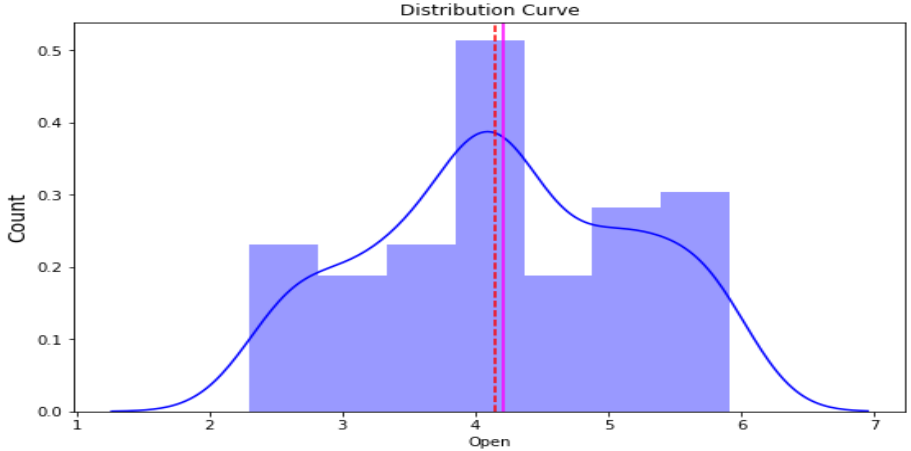
- The above graph shows that it is not a normal distribution curve.
- The mean and median should be equal for perfect normal distribution curve.But,mean is not equal to median as there is not a perfect normal distribution curve.
- We need to convert all the features to normal distribution using log transformation.
- Outliers are present in each column.By,converting our features to normal distribution using log transform We can remove outliers from the dataset.

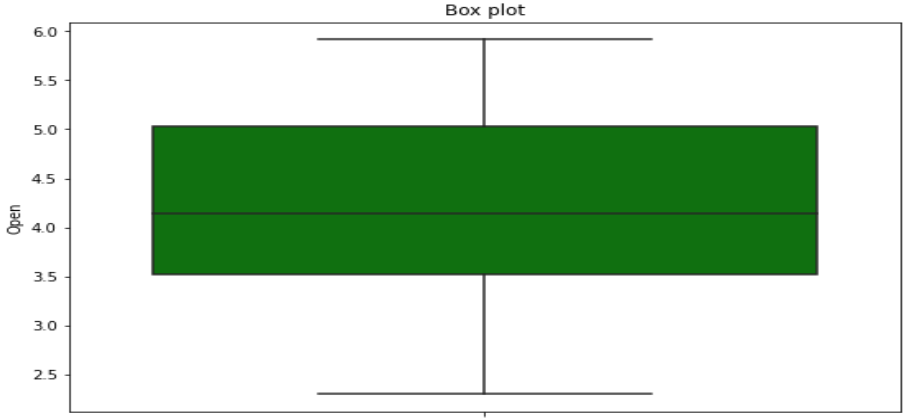
Transformation of Data

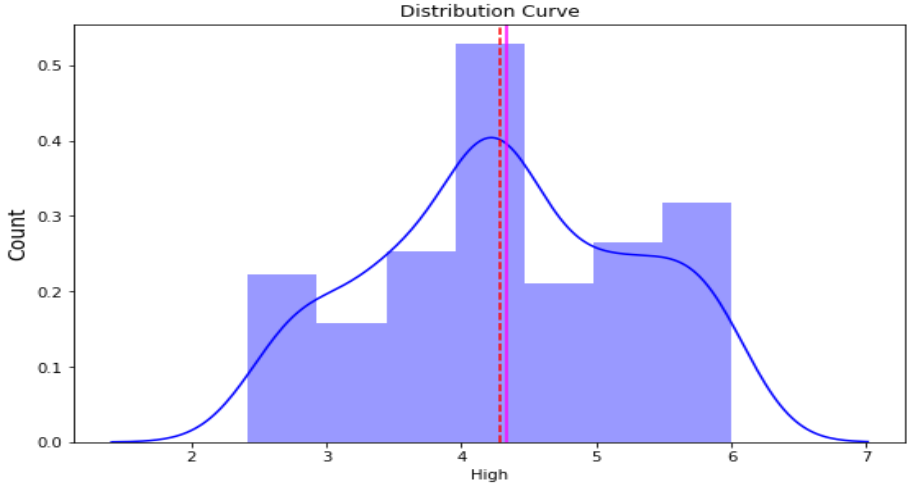
Data Transformation

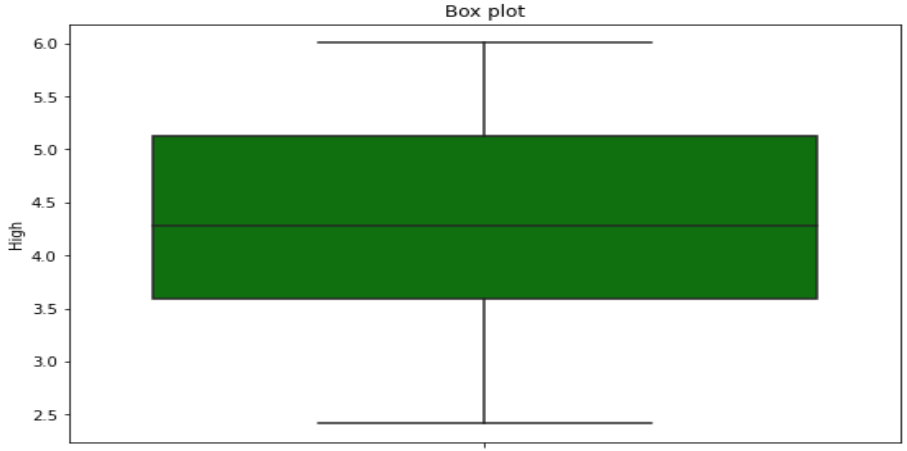
- As observed in the preceding slides, the observed data was found to be skewed
- We will transform the data to make it uniform before passing it into our ML models.
- Let's have a look at how they will look once the transformation is applied to them.
- The image on the right shows how the distribution of our close price would look after a log transformation is applied to it.

Now, You can see graph is normally distributed and outliers are removed

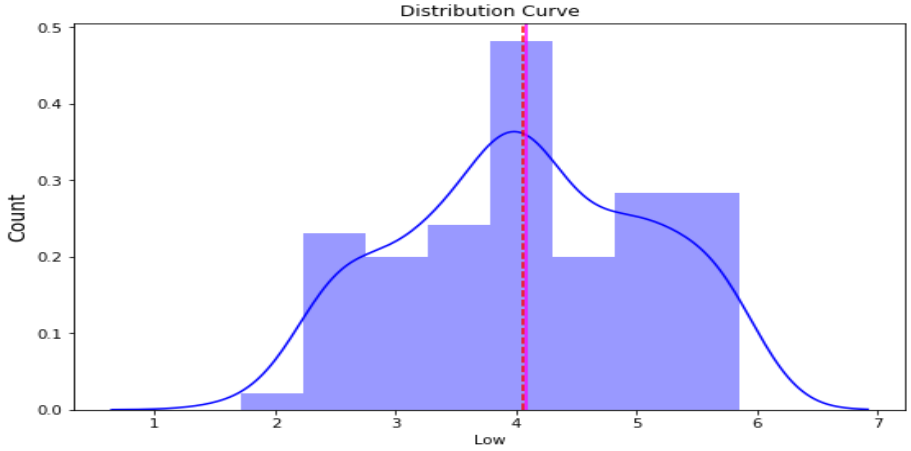


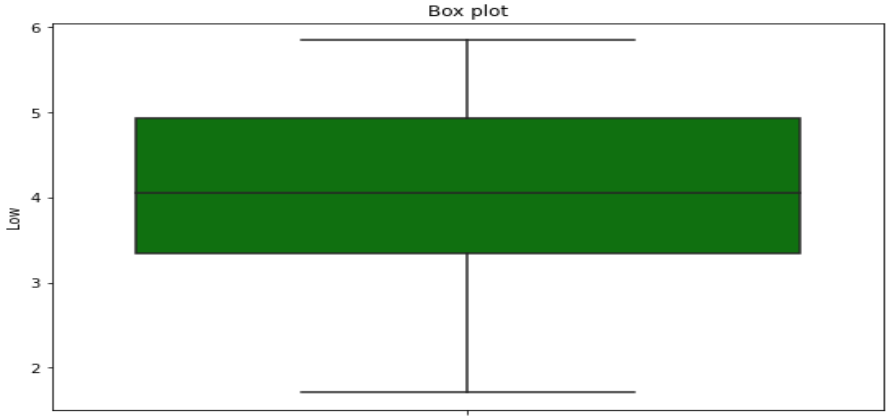


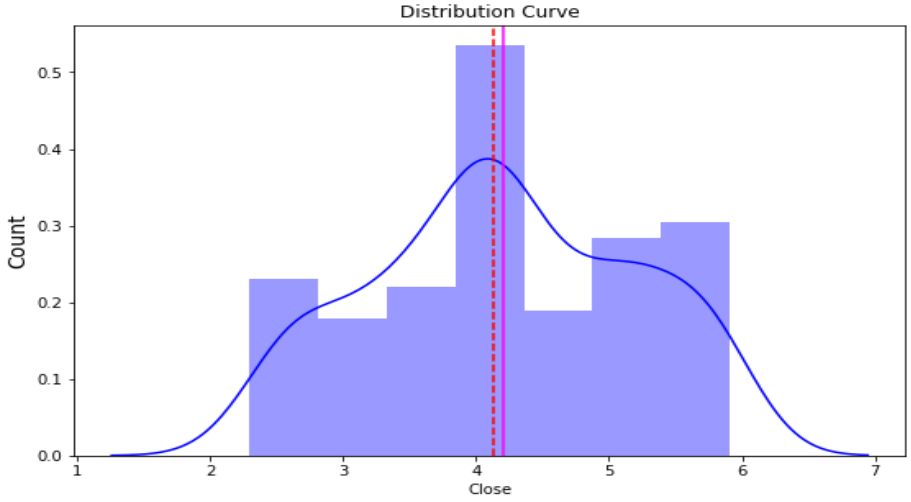


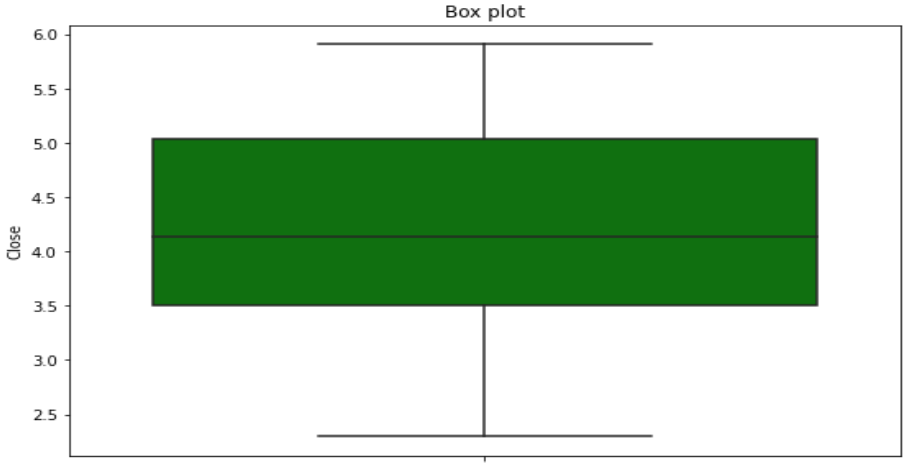


Now, You can see graph is normally distributed and outliers are removed

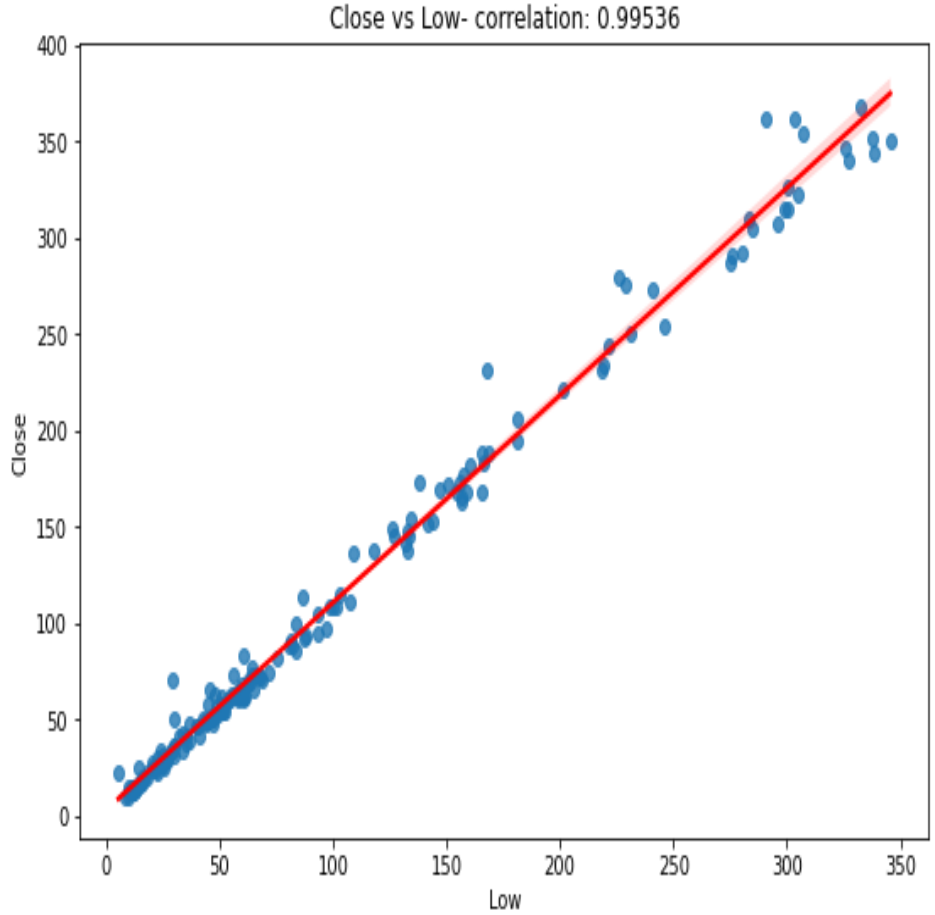


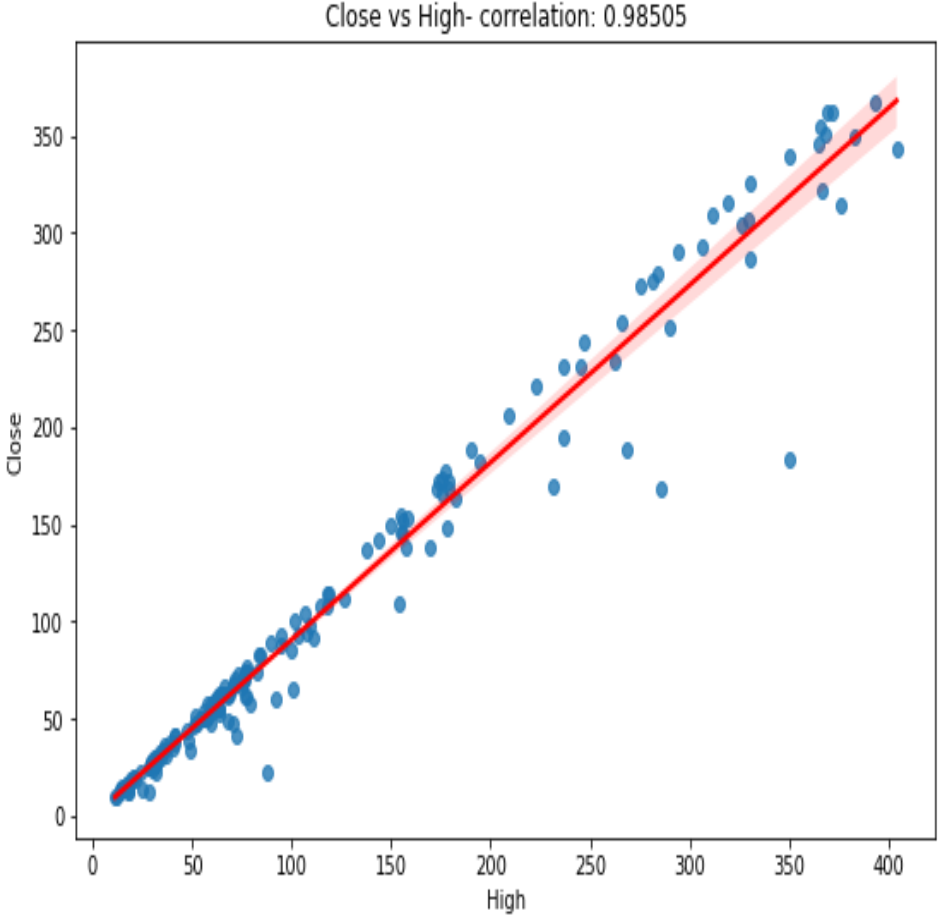






Correlation of ‘Closing Price’ with Independent Features:

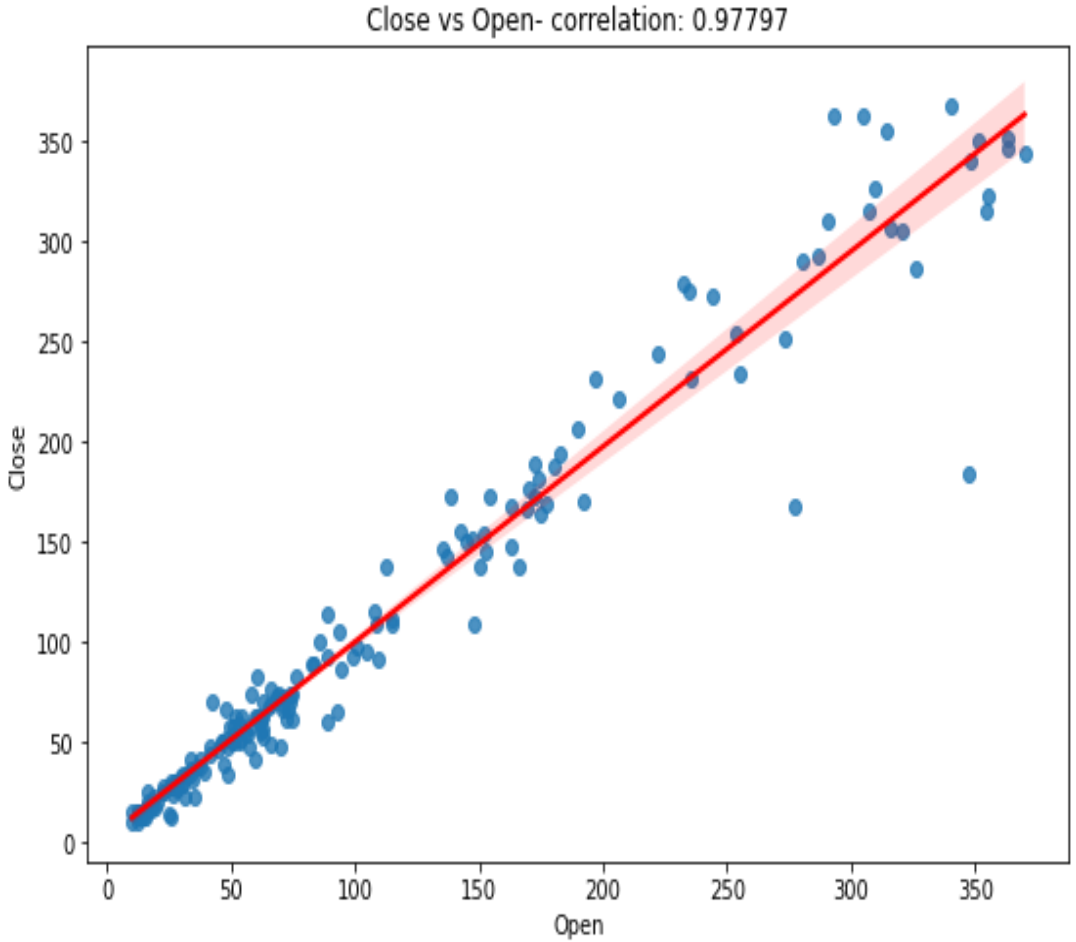




(Cont..)

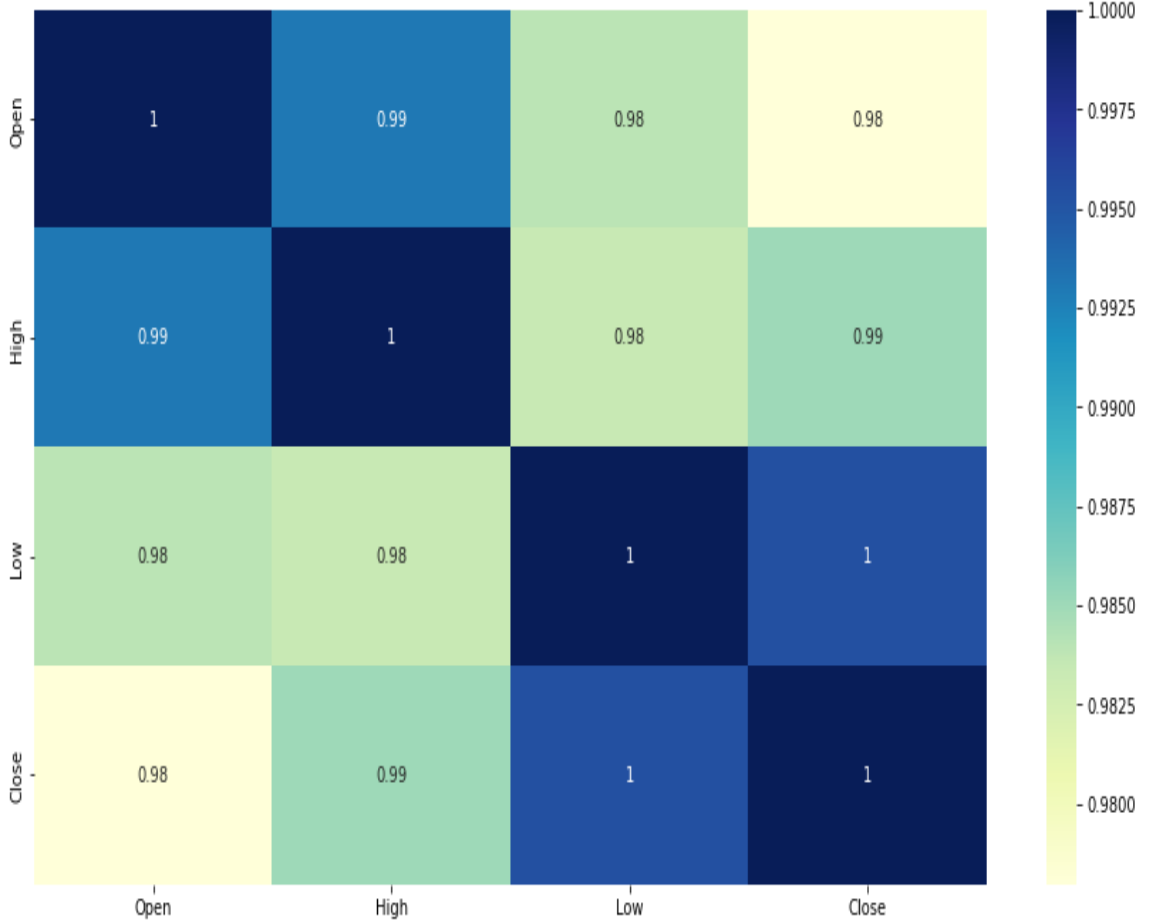
As we can see that there is linear relation and high correlation between each independent variables and dependent variable.

The correlation is 0.985, 0.995, 0.977 This suggests a high level of correlation, e.g. a value above 0.5 and close to 1.0.



Heat Map :showing relation between different features

- The Heat Map helps us visualize the correlation of each parameter with respect to every other parameter.
- The shades changes from the highest to lowest (or vice versa) correlations.
- We can see in the matrix on this slide that our dependent variable (close price) is highly correlated with all the other independent variables.



Model Selection

- We passed the data into different models like:
- Linear Regression
- Lasso Regression with and without Cross Validation
- Ridge Regression with and without Cross Validation
- Elastic Net with Cross Validation
- KNeighbor Regressor
- Support Vector Regressor

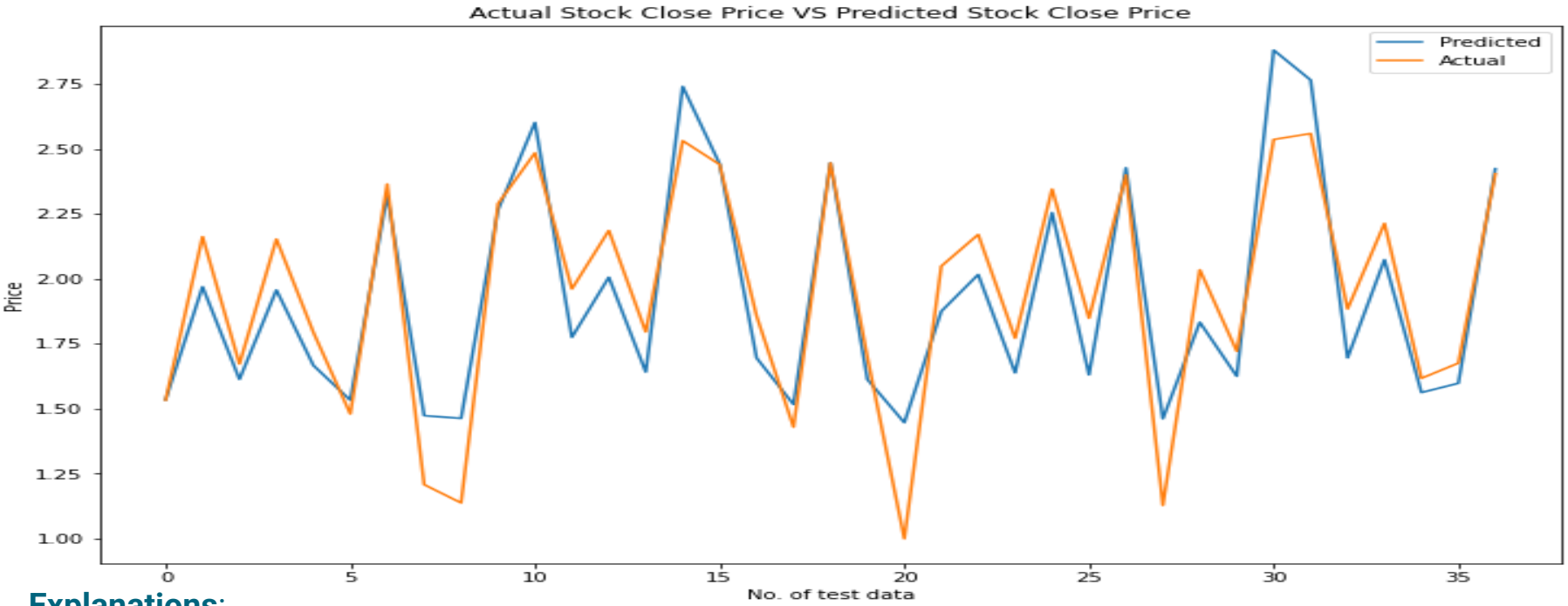
- We checked the performance of the each model across various parameters.
- Then we decided our best models on the basis of following metrics
- R2
- Adjusted R2
- Mean squared error and Root mean squared error.
- Mean absolute percentage error

Linear Regression

Explanations:

- Our Linear Model predicted the close price with 0.032 Mean Absolute error.
- R2 tells us that our independent is able to describe 83% of our dependent variable.
- Adjusted R2 is about 81.27%
- Mean Absolute Percentage Error is around 0.08%

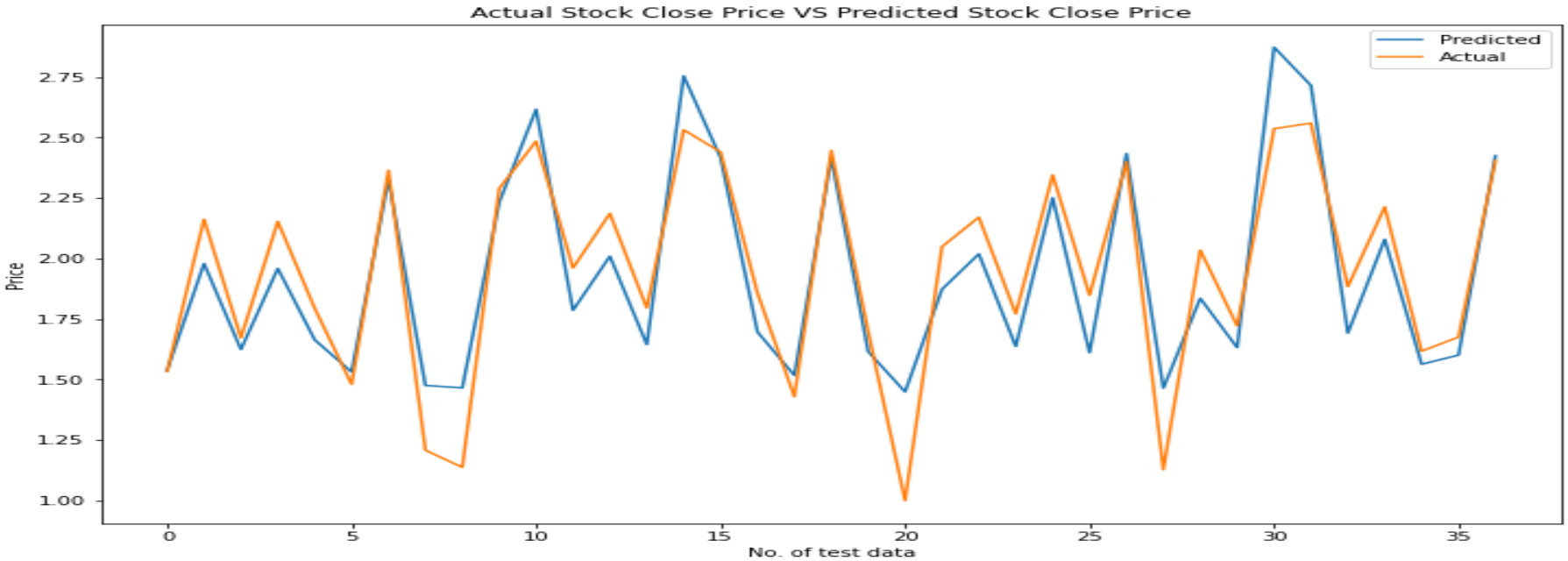
Ridge Regression



Explanations:

- Our Ridge predicted the close price with 0.03 Mean Absolute error.
- Here, R2 is about 0.8287 which means model's independent features is able to describe 82.8% of our dependent variable.
- Adjusted R2 is about 81.3%.
- Mean Absolute Percentage Error is 0.086%

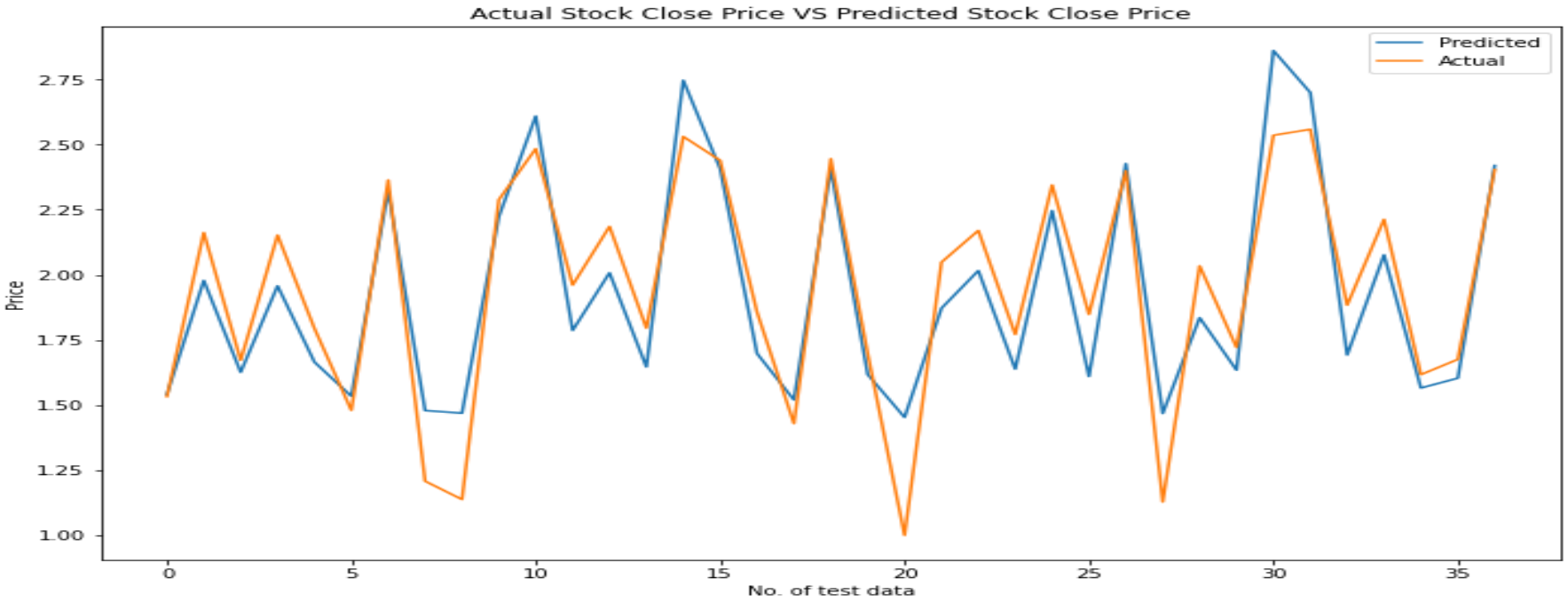
Ridge Regression with cross validation



Explanations:

- Our Ridge with CV predicted the close price with 0.03 1 Mean Absolute error.
- Here, R2 is about 0.8297 which means model's independent features is able to describe 82.97% of our dependent variable.
- Adjusted R2 is about 81.42% and Mean Absolute Percentage Error is 0.0874%
- It means there is no change in prediction even after using cross validation(CV).

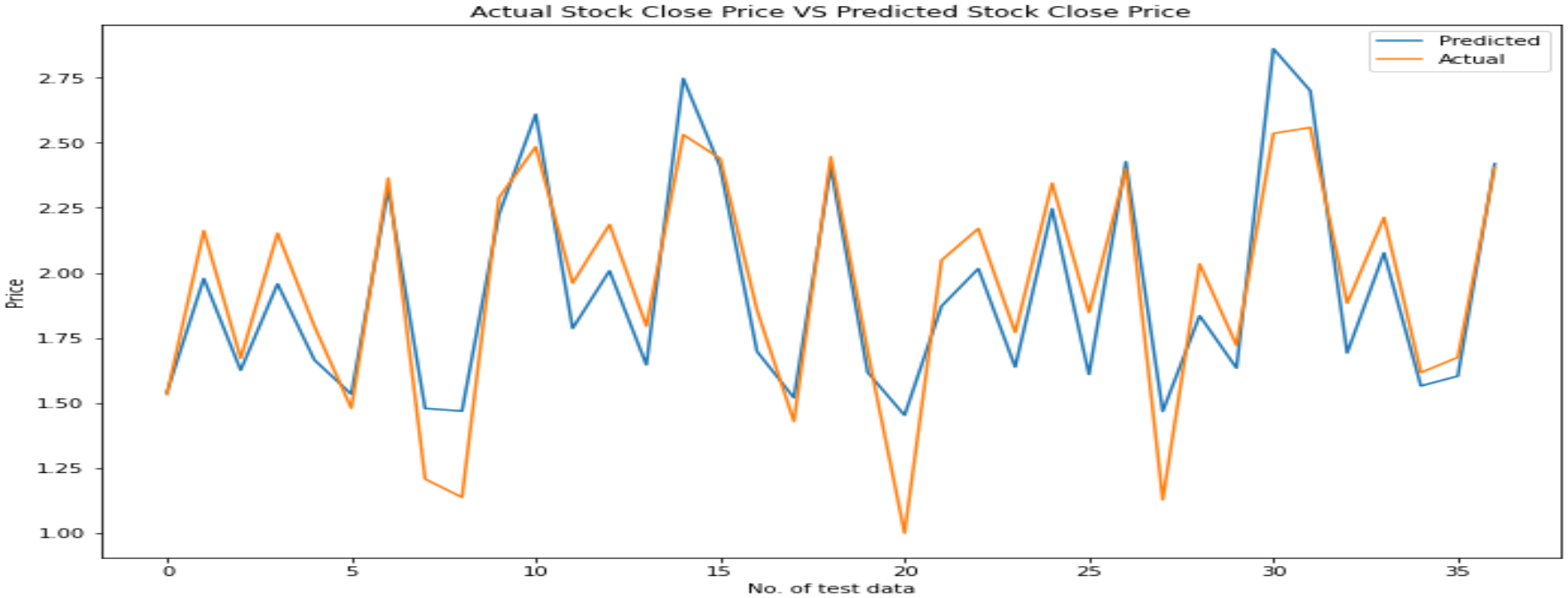
Lasso Regression



Explanations:

- Lasso predicted the close price with 0.031 Mean Absolute error.
- Here, R2 is about 0.8302 which means models' independent features is able to describe 83.02% of our dependent variable.
- Adjusted R2 is about 81.48%.
- Mean Absolute Percentage Error is 0.087%

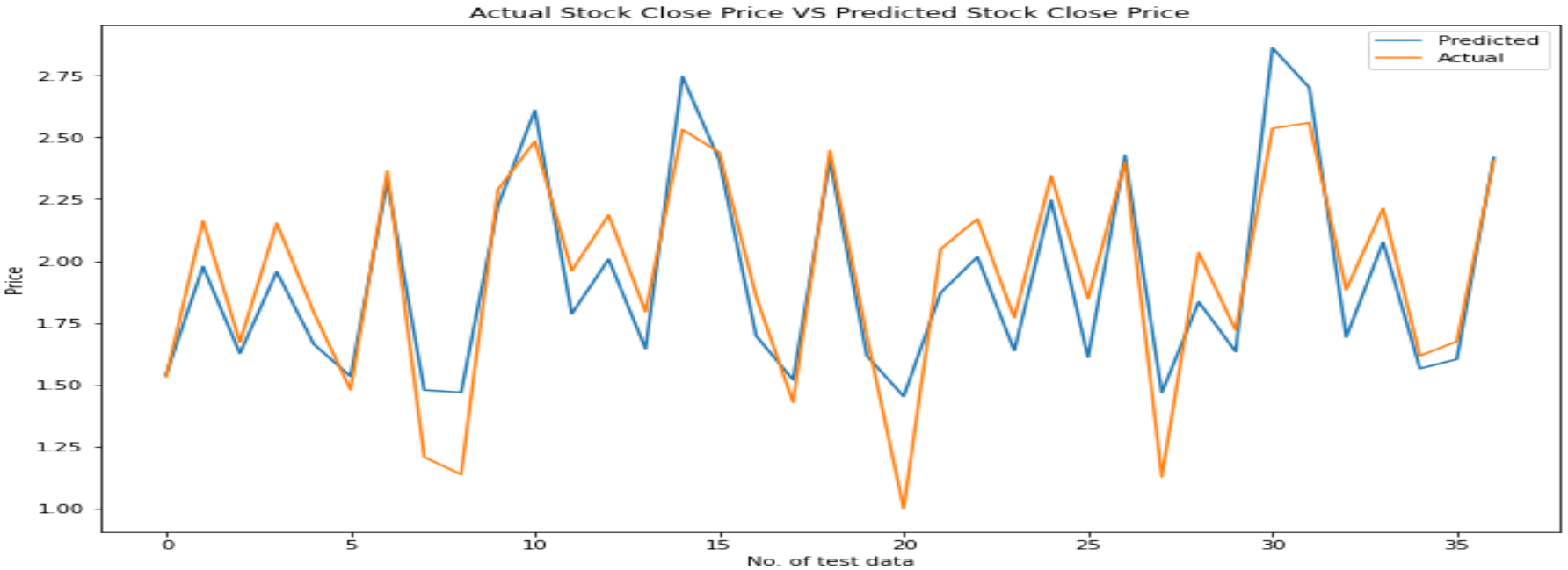
Lasso Regression with cross validation



Explanations:

- Lasso with CV predicted the close price with 0.031 Mean Absolute error.
- Here, R2 is about 0.8296 which means models' independent features is able to describe 82.96 % of our dependent variable.
- Adjusted R2 is about 81.41% and Mean Absolute Percentage Error is 0.0875%
- Very slight change in prediction after using CV.

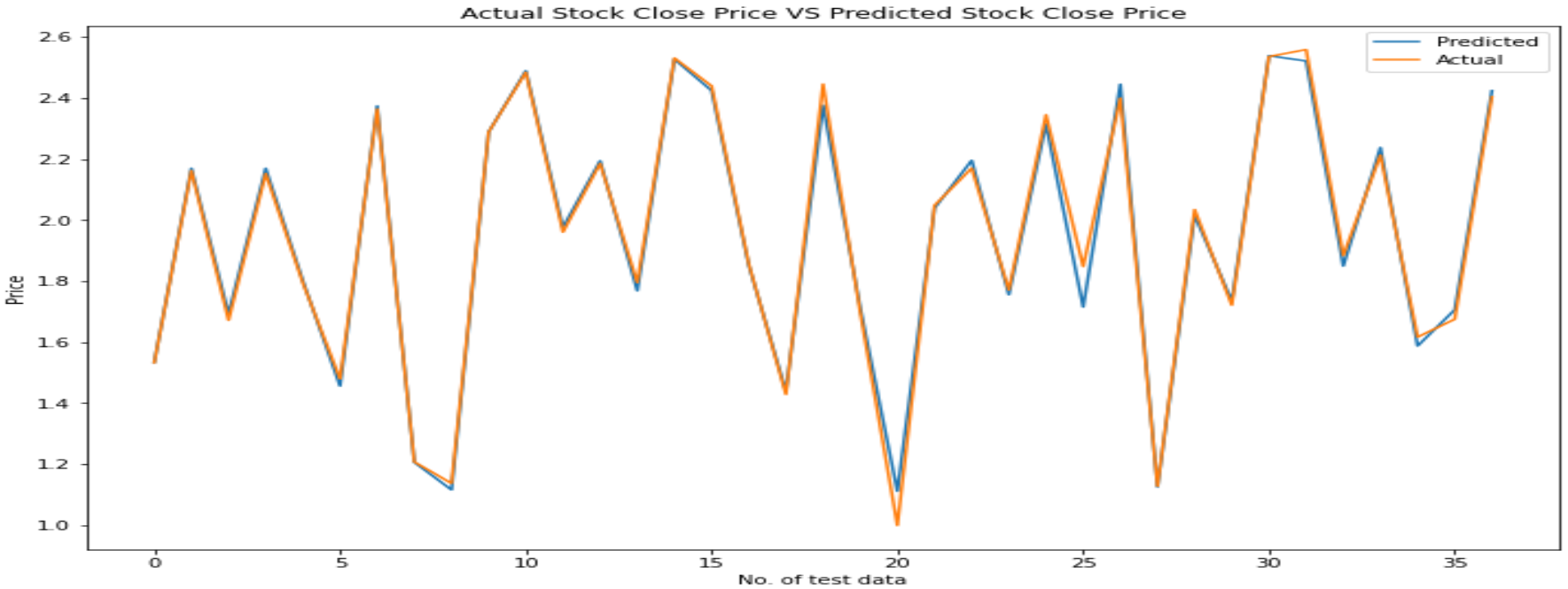
Elastic Net Regression Before Cross Validation



Explanations:

- Our elastic net Model predicted the close price with 0.0315 Mean Absolute error.
- R2 tells us that our independent is able to describe 83.03% of our dependent variable.
- Adjusted R2 is about 81.49%
- Mean Absolute Percentage Error is around 0.0876%

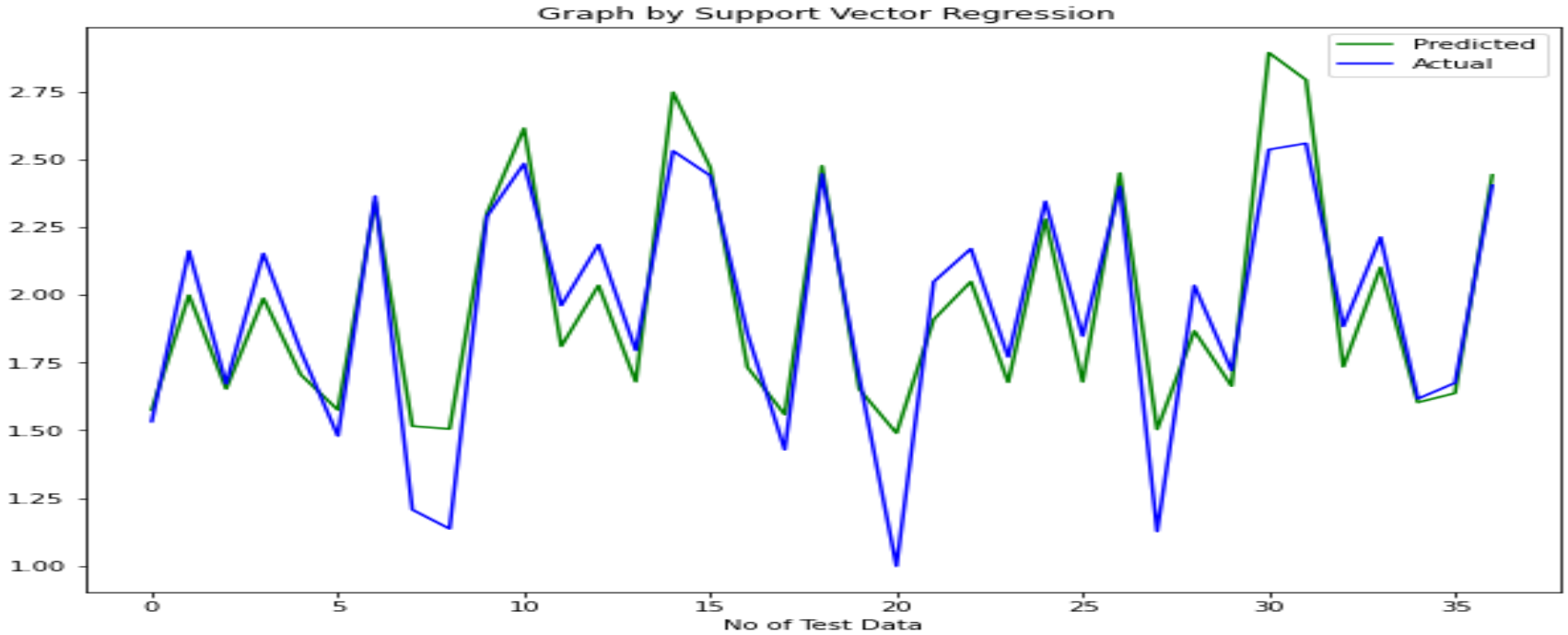
KNeighbor Regressor



Explanations:

- Our KNR Model predicted the close price with 0.0013 Mean Absolute error.
- R2 tells us that our independent is able to describe 99.29% of our dependent variable.
- Adjusted R2 is about 99.22% and Mean Absolute Percentage Error is around 0.0136%
- It performs best among all algorithms.

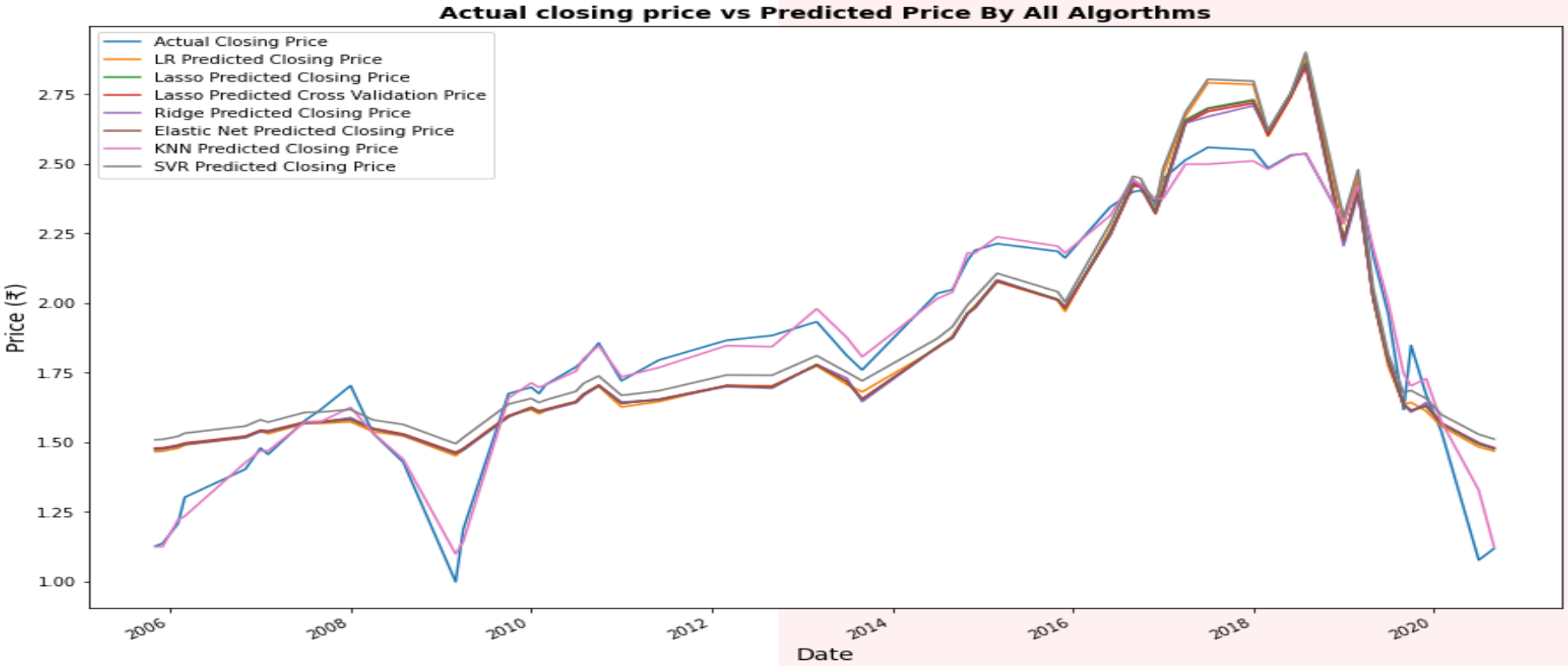
Support Vector Regressor



Explanations:

- Our SVR Model predicted the close price with 0.0317 Mean Absolute error.
- R2 tells us that our independent is able to describe 82.95% of our dependent variable.
- Adjusted R2 is about 81.40%
- Mean Absolute Percentage Error is around 0.0844%

Comparison among all Model predictions in one graph:



- KNR (KNeighbor Regressor) performs best.
- All models of linear regression i.e. (Ridge ,Lasso ,Elastic Net , Linear) gives nearly the same prediction.
- SVR performs low as compared to other algorithms.

Final Matrices for all models:

	Model_Name	MSE	RMSE	R2	Adjusted R2	MAPE
0	Linear regression	0.0326	0.1806	0.8310	0.8213	0.0918
1	Lasso regression	0.0321	0.1793	0.8334	0.8238	0.0922
2	Lasso Regression CV	0.0321	0.1792	0.8335	0.8239	0.0923
3	Ridge Regression	0.0325	0.1802	0.8316	0.8219	0.0917
4	Ridge Regression CV	0.0320	0.1789	0.8341	0.8245	0.0922
5	Elastiv Net CV	0.0321	0.1792	0.8336	0.8240	0.0922
7	KNeighbour Regressor	0.0029	0.0539	0.9850	0.9841	0.0213
8	Support Vector Regressor	0.0342	0.1850	0.8226	0.8123	0.0921

Conclusion

- Target Variable is strongly dependent on Independent Variables.
- We have seen that there is neither null nor duplicate values.
- But Date feature have values of object data type. So, We converted it into proper date format YYYY-MM- DD.
- KNeighbor Regressor and KNeighbor Regressor CV performing better than other models with adjusted R^2 0.9841 and 0.9916 respectively.
- With the help of visualization ,We have seen that from 2018 onwards there is sudden fall in the stock closing price. It makes sense how severely Rana Kapoor case fraud affected the price of Yes bank stocks.
- With the help of distribution plot ,We see that our data is positively skewed. So we apply some kind of transformation i.e. Log Transformation to convert it into Normal distribution.

Conclusion

- Lasso and Ridge regression models are giving the same result approximately.
- We have implemented Cross Validation on different algorithm as CV performs better on small datasets. But, the result is nearly same.
- In all the models except KNeighbor Regressor, the accuracy lie within the range of 81 to 83% and there is no such improvement in accuracy score even after hyperparameter tuning.
- Support Vector Regressor algorithm performs worst then other algorithm with accuracy of 81.2 % .
- KNR cross validation perform best with very less mean square error i.e. 0.015

Thank you ....